DP-FedLoRA: Privacy-Enhanced Federated Fine-Tuning for On-Device Large Language Models

Honghui Xu[1], Shiva Shrestha[1], Wei Chen[2], Zhiyuan Li[2], Zhipeng Cai[3]

Kennesaw State University[1], Nexa AI[2], Georgia State University[3]

Introduction

- Large Language Models (LLMs) are now deployed on edge devices (like smartphones) for personalized AI.
- Federated Learning (FL) is used to fine-tune these models on private user data without centralizing it.
- This process, while protecting raw data, still creates a significant privacy risk.

The Threat: MIA

- The primary threat is the Membership Inference Attack (MIA).
- A semi-honest server can analyze the model updates to infer if a specific user's private data was used during training, leaking sensitive information- thus motivating our research.

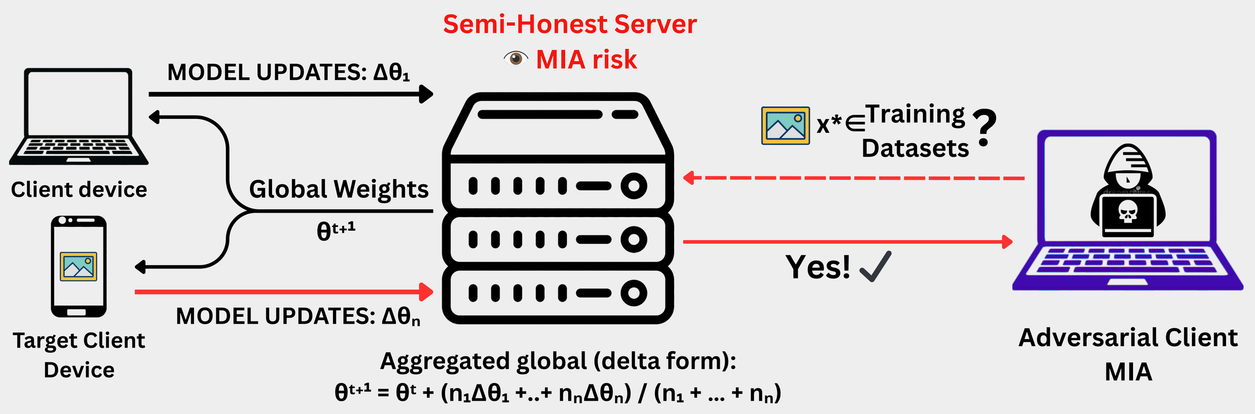

Fig. 1: Membership inference in a semi-honest server setting.

Methodology

DP-FedLoRA Algorithm

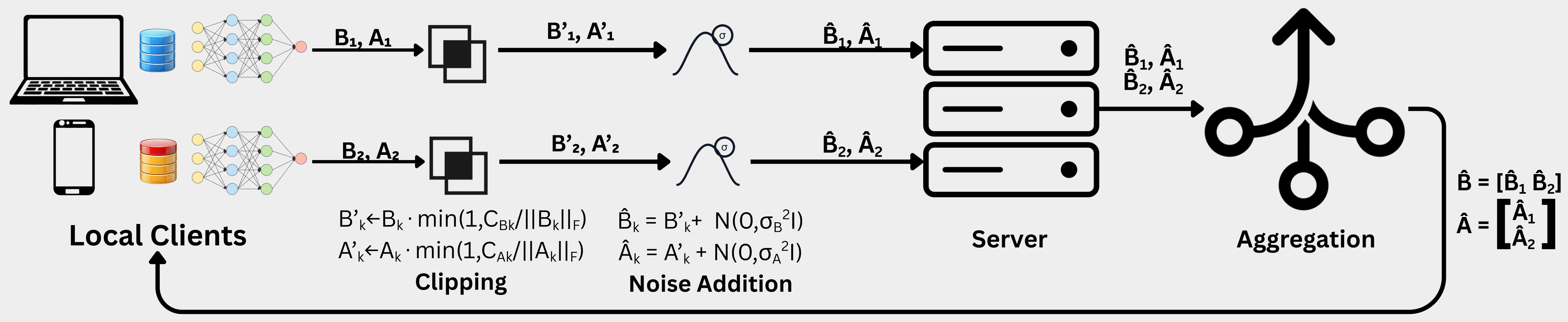

Fig. 2: DP-FedLoRA Flowchart

1. **Local Training:** Each client 'K' fine-tunes its LoRA matrices (B_k, A_k) on its private data.
2. **Privacy Injection:** Before sending, each client 'k' clips the norm of its matrices (B_k, A_k) and adds calibrated Gaussian noise to create new private update($\hat{B}_k$, $\hat{A}_k$).
3. **Secure Aggregation:** The server performs structured stacking to create a global, private update ($\hat{B}$, $\hat{A}$) and broadcasts it.

Theoretical Guarantees

Our privacy mechanism is supported by key theoretical guarantees:

1. **Unbiased Updates:** The noise is centered at zero and doesn't systematically skew the model, ensuring it converges correctly on average. $E[\hat{B}\hat{A}]-E[BA]=0$
2. **Bounded Variance:** We provide an analytical bound on the variance (the "spread" of error) caused by the noise, allowing us to manage the privacy-utility trade-off.

$$Var[\hat{B}\hat{A}] \leq m\sigma_\alpha^2\|B\|_F^2 + n\sigma_\beta^2\|A\|_F^2 + \sigma_\alpha^2\,\sigma_\beta^2.m.n.r$$

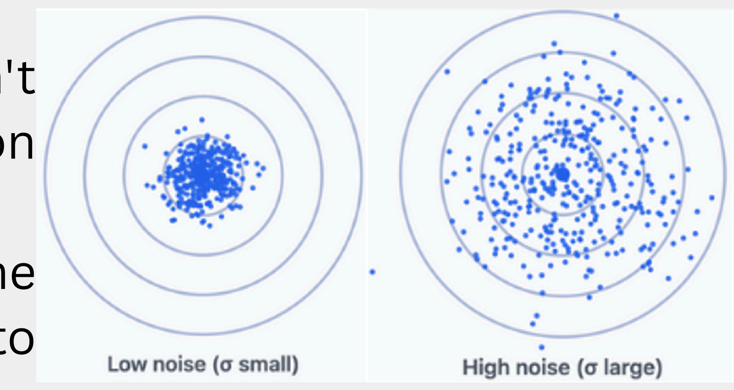

Fig. 3: A conceptual analogy for our theoretical guarantees

Results

Benchmark Performance

We tested DP-FedLoRA (ϵ = 25.0 and cliping norm of 0.1) against non-private baselines using a LLaMA-2-7B model.

- **Knowledge & Reasoning are Robust**: On MMLU (knowledge) and BBH (reasoning) benchmarks, the performance drop from adding privacy is very similar and minimal, averaging only 4-5%.
- **Counterfactual Reasoning is Sensitive**: The CRASS benchmark showed a more significant performance drop, proving more sensitive to privacy noise.

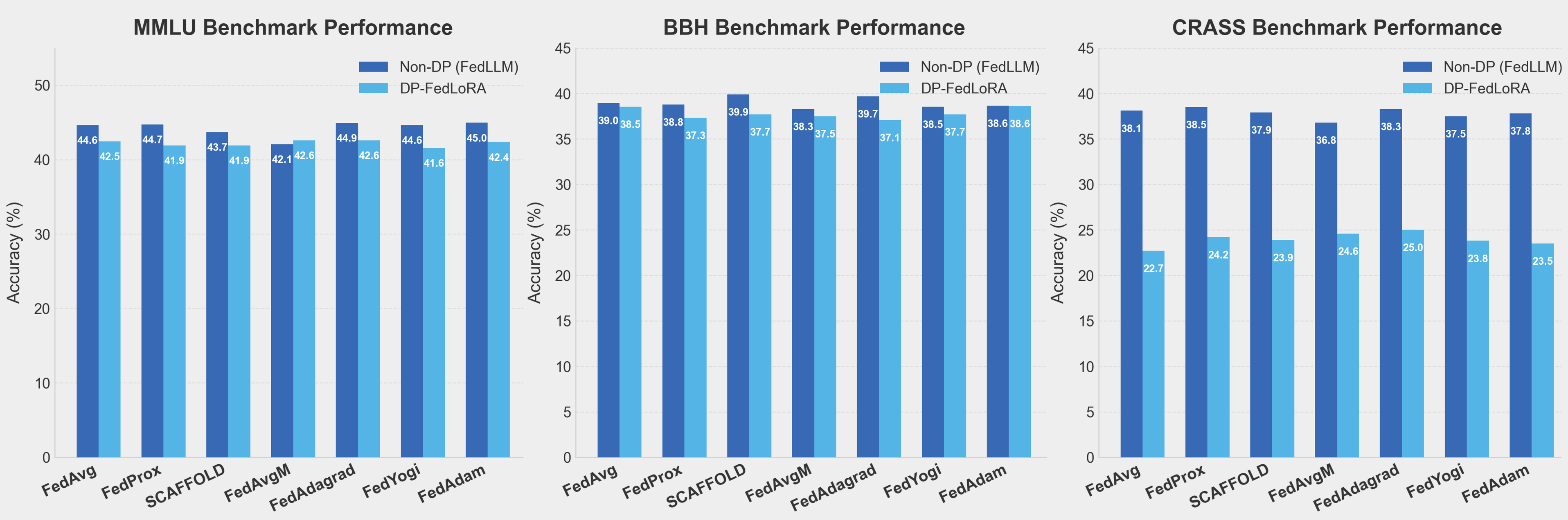

Fig. 4: Benchmark Performance of DP-FedLoRA (DP) vs. Non-Private (Non-DP) Baselines

Validating Theory

Our experiments empirically confirm our theoretical guarantees from the "Methodology" section:

- **Unbiased Expectation Confirmed:** Across all tests, the expected difference between the noisy/private update and non-private update remained centered at zero.
- **Variance Scales as Predicted:** The variance of the update (the "spread") grew linearly as the LoRA rank increased and the number of parameters of the LLM increased.

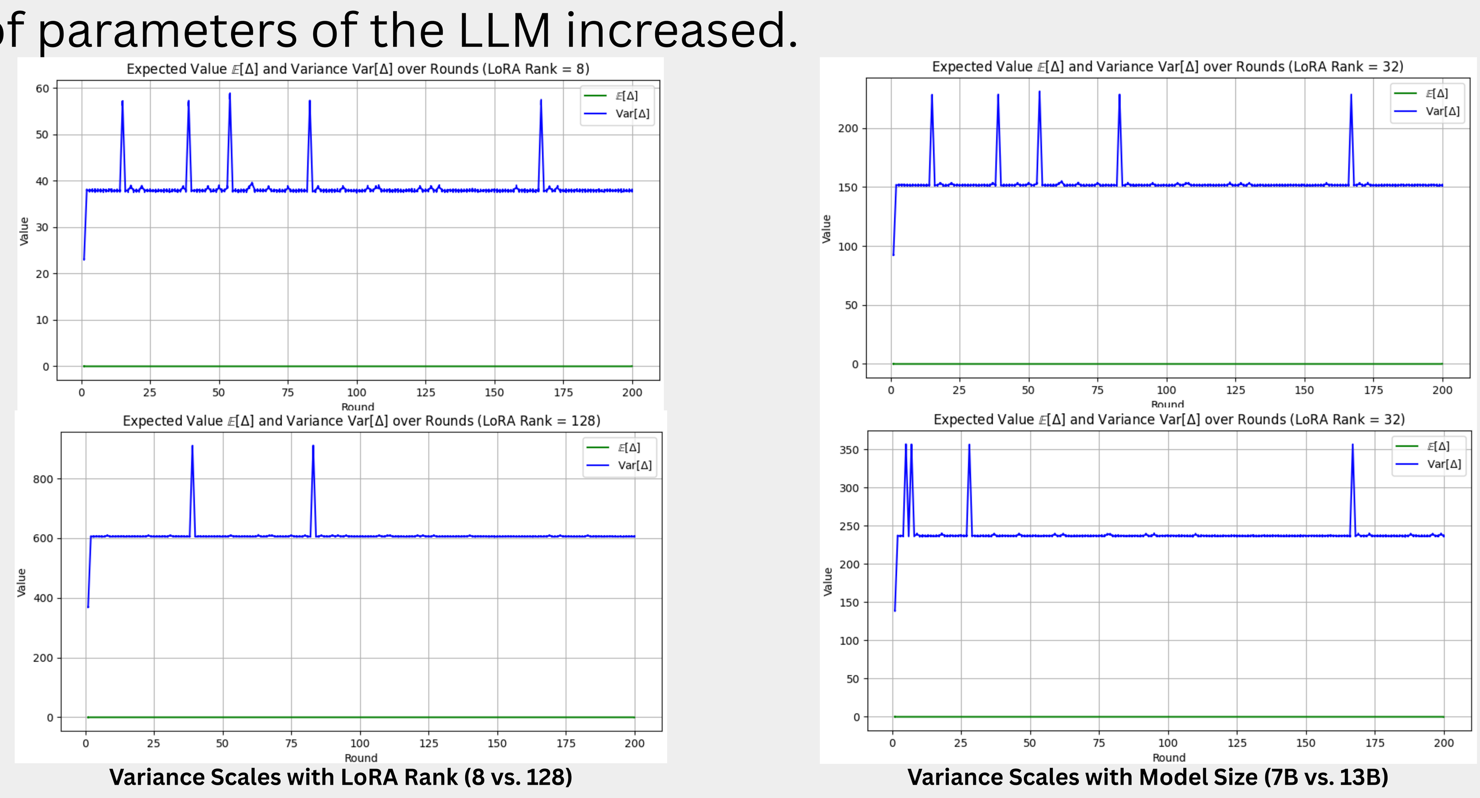

Fig. 5: Empirical Validation of Theoretical Guarantees: Expectation and Variance.

Conclusion

- **Strong Privacy with Minimal Performance Loss:** Our framework DP-FedLoRA introduces differential privacy into federated LoRA fine-tuning, achieving <5% average accuracy drop across MMLU and BBH benchmarks.
- **Algorithm-Level Robustness:** Consistent results across FedAvg, FedProx, and adaptive optimizers demonstrate generality for on-device LLM personalization.
- **Practical Edge Deployment:** The framework supports efficient, private, and communication-aware fine-tuning on resource-constrained edge devices.

References

- E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen et al., "Lora: Low-rank adaptation of large language models." ICLR, vol. 1, no. 2, p. 3, 2022.
- T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," 2020. [Online]. Available: https://arxiv.org/abs/1812.06127
- Ye, R., Wang, W., Chai, J., Li, D., Li, Z., Xu, Y., Du, Y., Wang, Y., & Chen, S. (2024). OpenFedLLM: Training Large Language Models on Decentralized Private Data via Federated Learning. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- Z. Wang, Z. Shen, Y. He, G. Sun, H. Wang, L. Lyu, and A. Li, "Flora: Federated fine-tuning large language models with heterogeneous lowrank adaptations," arXiv preprint arXiv:2409.05976, 2024.

Resources & Contact

Paper GitHub

Intelli-Trust Lab

Paper ID: DM489